SOME FACTORS INFLUENCING THE UNRELIABILITY OF CLINICAL JUDGMENTS *

Franklyn N. Arnhoff, Ph.D.
Northwestern University and Downey V.A. Hospital

The unreliability of clinical judgment is well known, particularly in the field of diagnosis (1, 8). Little has been done however, toward any analysis of the factors responsible for this unreliability in order that it may be understood, controlled and corrected. Magaret (7), recently has pointed out that we need both a philosophy of diagnosis and a sophisticated understanding of its nature. Hunt, Wittson, and Hunt (4, 5), have suggested that our understanding of clinical judgment might be furthered if we conceived of it, not as a unique and special kind of professional performance, but as one example of the broader phenomenon of human judgment in general. The present study takes this approach, and studies the effect upon the clinical judgment of both the professional experience of the judge, and of anchoring the scale which he us using to make his judgment. Both experience (2, 10) and anchoring (3, 9) have been shown to influence judgment in a wide range of situations varying from those of classical psychophysics to the judgments of the prestige of occupations and the undesirability of certain forms of behavior.

STATEMENT OF PROBLEM

We assumed that clinical judgments might show the same relativism that has been demonstrated in other fields of judgment and that this relativism might be contributing to the unreliability of the clinical judgments. Specifically we proposed three hypotheses:

(1) That introducing an anchoring stimulus at either end of the stimulus continuum would cause a shift in the judged value of the stimuli being evaluated;

---

* This paper is a condensation of a longer one submitted to the Graduate School of Northwestern University in partial fulfillment of the requirements for the Ph.D. degree. The complete thesis, containing a more detailed statement of procedure and all the relevant data with their complete statistical analyses, is available from University Microfilms, Ann Arbor, Michigan. The study is part of a larger project being conducted at Northwestern University under Professor William A. Hunt through contract 7onr-45011 with the Office of Naval Research. The opinions expressed, however, are those of the author and do not represent the opinions or policy of the Naval service. Thanks are due Professor B. J. Underwood for assistance with the design, and Dr. J. W. Cotton for assistance with the statistical analyses.

(2) That these anchoring effects would be a function of the experience of the judges, with the most experienced judges showing the least shift; and

(3) That the reliability of the judgments, here defined as inter-judge agreement, would also be a function of experience, with the most experienced judges showing the greatest reliability or agreement.

SUBJECTS AND MATERIALS

In order to test the effect of experience, three groups of judges were selected from three separate levels of clinical training. Sixty were undergraduates who had just completed a course in abnormal psychology; sixty were graduate students interning during a clinical psychology training program, and sixty were professional clinicians with four years or more on-the-job professional experience. As stimuli, schizophrenic responses to items on the Wechsler-Bellevue and Terman-Binet vocabulary tests were used and the subjects were asked to rate these on an 11-point scale for the severity of the disorder in the thinking processes exhibited in the responses.

PROCEDURE

For the construction and equation of the two stimulus series used in the experiment, it was first necessary to obtain a number of stimuli whose stimulus values were known. 222 schizophrenic responses to vocabulary items, judged by a group of 3 trained clinicians to cover all possible values of confusion in thinking in such responses, were rated on an 11-point scale by another 22 experienced clinicians of at least four years professional experience. These judges were not used in the experiment proper. They merely served as a standardization group. The means and standard deviations of the stimuli were then computed. This furnished a group of standardized stimuli from which two roughly equivalent stimulus series were constructed. Each series contained 10 items - 2 of each representing scale values 4, 5, 6, 7 and 8. The stimuli thus represented only the middle ranges of the continuum of "confusion" since scale values 1, 2, 3, 9, 10, and 11 were not represented in the series. It was felt that the use of a limited range of stimuli would offer more room for movement or shifts under the anchoring conditions.

All sixty subjects at each of the three experience levels were given the first series with the same instructions used for the original standardization group of 22 clinicians. Following this, each group of sixty subjects was split into three sub-groups of 20 each. Each of these sub-groups then were presented with the second series of stimuli. The first sub-group received the previous instructions but with an anchor at the high end of the scale. This was done by adding the following to the standard instructions: "In order to further assist you in defining the scale, we will give you the following as an illustration of a response which represents the category eleven: FABLE: Trade good sheep to hide in the beginning." The second group received the stimuli with an anchor at the low end of the scale as follows: "In order to further assist you in defining the scale, we will give you the following as an illustration of a response

which represents the category one: GAMBLE: To take a chance, a risk." The third group served as a control and got the second series with no anchor and no change in instructions.

In processing the data the mean and standard deviation of each subject's ratings of each stimulus series was computed. These obtained means and standard deviations then were themselves treated as though they were raw scores and a mean and standard deviation for the distribution of means and the distribution of standard deviations were computed. This was done with both stimulus series for each experience level of 60 subjects, and for each sub-group of 20 within the experience levels. The combining of the three sub-groups at each experience level on the second stimulus series was justified by the fact that no significant effects for the anchoring conditions were found within any experience level.

Comparisons of results between sub-groups within an experience level as well as between experience levels on a single scale were made by analyses of variance. Overall comparisons for all sub-groups and experience levels on both scales, were made by analyses of covariance.[1] Bartlett's Chi Square tests were used for determinations of homogeneity of variance. As measures of reliability of the judges' ratings, the standard deviations were used as primary measures, and were supplemented by an r which is recommended by Johnson (6, p. 134).

RESULTS

Analysis of covariance of the mean ratings for the various sub-groups (anchoring conditions) and the three experience levels failed to demonstrate significance, indicating that the introduction of the anchoring stimuli failed to produce changes that were any greater or less than those occuring in the control groups. This was found to hold for all three experience levels. Thus our first hypothesis was not substantiated.

Since no anchoring effects appeared, our second hypothesis (that anchoring effects would be a function of experience) becomes meaningless.

To test our hypothesis concerning reliability we compared the variances of the mean ratings of the judgments at each of the three experience levels on the first stimulus series. The data are presented in Table I. There were no significant differences between the means. Comparison of the variances of these means by a Bartlett's test for homogeneity of variance, however, yielded a X^2 of 16.94 for 2 d.f., which is significant beyond the .01 level. While this finding supports our hypothesis regarding differences in reliability (inter-judge agreement) between the three

---

[1] In the interests of brevity, only the results of our statistical analyses are included here. The complete statistical treatment of data is available on microfilm as mentioned in the introductory note.

TABLE 1

Means and Standard Deviation for Experience Levels
Based upon Mean Rating by Individual Judges

	Mean	Standard Deviation
Scale 1		
Clinicians	5.86	1.28
Trainees	5.84	.99
Students	6.15	.74
Scale 2		
Clinicians	6.30	1.25
Trainees	6.51	1.34
Students	6.41	.94

experience levels, the results are a complete reversal of the predicted direction. Our professional clinicians are least reliable, our trainees next, and the undergraduates most reliable.

As stated above, the absence of any demonstrable effect from our anchoring conditions enabled us to combine the sub-groups of 20 at each experience level for stimulus series 2 and treat the data for reliability as done with stimulus series 1. These data also are found in Table I. Again there were no significant differences between the means, but a Bartlett's test comparing the variance of the means gave a X^2 of 7.65 for 2 d.f., significant at the 5% level. This time, however, professional clinicians and trainees reversed positions. The trainees were least reliable, the professional clinicians next, and the undergraduates again most reliable.

As mentioned before correlations were computed using Johnson's formula (6, p. 134) as a further measure of the reliability of the judgments. While there is no adequate method for evaluating the significance of the differences between the r's obtained, inspection shows the previous findings to be confirmed. The undergraduates showed the greatest inter-judge agreement, the trainees less, and the professional clinicians least.

DISCUSSION

Within the limits of this experiment and for this type of judgment professional experience and training would appear to result in lowered reliability. At first glance, this might seem to be a disastrous reflection upon clinical training. Upon further consideration, however, the results are quite understandable in the light of the increased possibility of differing self-instructions, differing interpretations of the standard instructions, etc. for our experienced groups. It may well be that increased training and professional experience provides the experienced clinician with multiple frames of reference against which to evaluate behavior. These multiple frames of reference provide diverse grounds on which the actual judgment may be based, as was obvious from spontaneous comments offered by our experienced clinicians. Clang associations were sometimes viewed as not "severe," paranoid thinking was not considered "disordered" by some, and some subjects indicated that they had made their judgments not on the severity of the disorder exhibited, but on its indication for therapeutic accessibility or on its prognostic value for recovery.

There is perhaps, one homely caution that may be drawn from these results if our interpretation is correct. When dealing with experts in a judgmental situation, the task should be well defined and the criteria set forth clearly. Otherwise the riches of knowledge may yield confusion rather than clarity.

SUMMARY

Subjects with different degrees of professional clinical experience rated schizophrenic Wechsler-Bellevue and Terman vocabulary responses on an 11-point scale for degree of disorganization of thinking. Anchoring

values were introduced as a means of influencing the judgments made. Specific hypotheses were advanced regarding the effects of experience and anchoring upon the judgments made. No significant results due to anchoring could be demonstrated. Inter-judge agreement was found to decrease as a function of increasing experience.

REFERENCES

1. Asch, P. The reliability of psychiatric diagnoses. J. abnorm. soc. Psychol., 1949, 44, 272-276.

2. Doughty, J. M. The effect of psychophysical method and context on pitch and loudness functions. J. exper. Psychol., 1949, 39, 729-745.

3. Hunt, W. A. Anchoring effects in judgment. Amer. J. Psychol., 1941, 54, 395-403

4. Hunt, W. A., Wittson, C. L., and Hunt, E. B. A theoretical and practical analysis of the diagnostic process. In P. H. Hoch and J. Zubin (Eds.), Current problems in psychiatric diagnosis. New York: Grune and Stratton, 1953, pp. 53-65.

5. Hunt, W. A., Wittson, C. L., and Hunt, E. B. Relationship between definiteness of psychiatric diagnosis and severity of disability. J. clin. Psychol., 1952, 8, 314-315.

6. Johnson, P. O. Statistical methods in research. New York: Prentice-Hall, 1949.

7. Magaret, A. Clinical methods: psychodiagnostics. In Stone, C. P. and Taylor, D. W. (Eds.) Annual Review of Psychology 1952. Annual Reviews: Stanford, 1952, pp. 283-320.

8. Mehlman, G. The reliability of psychiatric diagnoses. J. abnorm. soc. Psychol., 1952, 47, 577-578.

9. Volkmann, J. Scales of judgment and their implications for social psychology. In J. H. Rohrer and M. Sherif (Eds.), Social psychology at the crossroads. New York: Harper Bros., 1951, pp. 273-294.

10. Wever, E. G., and Zener, K. E. Method of absolute judgment in psychophysics. Psychol. Rev., 1928, 35, 466-493